IDS

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06F 15/16, 17/00, 17/30 — A1

(11) Internati nal Publicati n Number: WO 00/20982

(43) International Publication Date: 13 April 2000 (13.04.00)

(21) International Application Number: PCT/US99/22966

(22) International Filing Date: 1 October 1999 (01.10.99)

(30) Priority Data:
60/102,831 2 October 1998 (02.10.98) US

(71) Applicant *(for all designated States except US)*: NCR CORPORATION [US/US]; 101 W. Schantz Avenue, Dayton, OH 45479 (US).

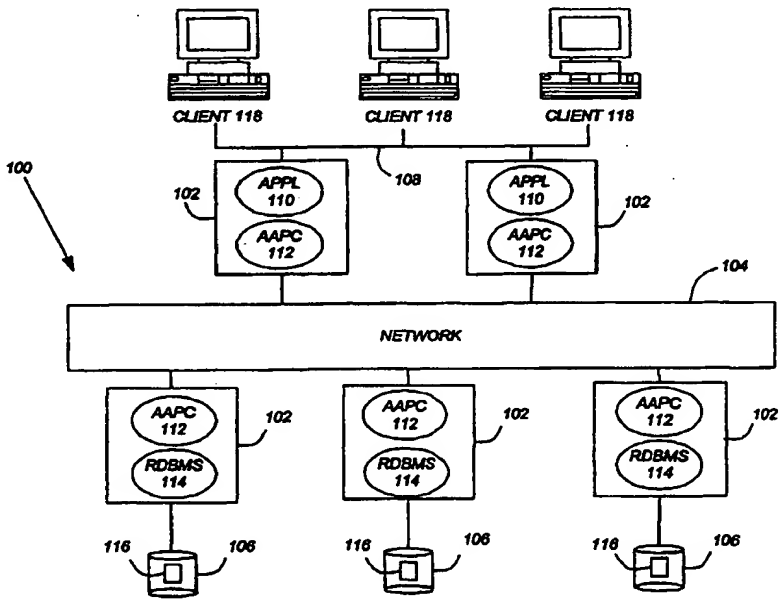
(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: MILLER, Timothy, Edward [US/US]; 32668 Hupa Drive, Temecula, CA 92592 (US). TATE, Brian, Don [US/US]; 314 Skyridge Lane, Escondido, CA 92026 (US). HILDRETH, James, Dean [US/US]; 1545 Chandelle Lane, Fallbrook, CA 92028 (US). BRYE, Todd, Michael [US/US]; 12387 Briardale Way, San Diego, CA 92128 (US). ROLLINS, Anthony, Lowell [US/US]; 12502 Pacato Circle South, San Diego, CA 92128 (US). PRICER, James, Edward [US/US]; 2614 Winningham Road, Chapel Hill, NC 27516 (US). ANAND, Tej [US/US]; 71 Pond View Lane, Chappaqua, NY 10514 (US).

(74) Agents: STOVER, James, M.; NCR Corporation, 101 W. Schantz Avenue, Dayton, OH 45479 (US) et al.

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: SQL-BASED ANALYTIC ALGORITHMS



(57) Abstract

A method, apparatus, and article of manufacture for performing data mining applications in a relational database management system. At least one analytic algorithm (110) is perf rmed by a computer directly against a relational database (116), wherein the analytic algorithm includes SQL statements performed by the relational database management system (114) and optional programmatic iteration, and the analytic algorithm creates at least one analytic model within an analytic logical data model from data residing in the relati nal database.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

SQL-BASED ANALYTIC ALGORITHMS

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims the benefit under 35 U.S.C. Section 119(e) of the co-pending and commonly-assigned U.S. provisional patent application Serial No. 60/102,831, filed October 2, 1998, by Timothy E. Miller, Brian D. Tate, James D. Hildreth, Miriam H. Herman, Todd M. Brye, and James E. Pricer, entitled Teradata Scalable Discovery, which application is incorporated by reference herein.

This application is also related to the following co-pending and commonly-assigned utility patent applications:

Application Serial No. --/ - - -, - - -, filed on same date herewith, by Brian D. Tate, James E. Pricer, Tej Anand, and Randy G. Kerber, entitled SQL-Based Analytic Algorithm for Association, attorney's docket number 8219,

Application Serial No. --/ - - -, - - -, filed on same date herewith, by James D. Hildreth, entitled SQL-Based Analytic Algorithm for Clustering, attorney's docket number 8220,

Application Serial No. --/ - - -, - - -, filed on same date herewith, by Todd M. Brye, entitled SQL-Based Analytic Algorithm for Rule Induction, attorney's docket number 8221,

Application Serial No. --/ - - -, - - -, filed on same date herewith, by Brian D. Tate, entitled SQL-Based Automated Histogram Bin Data Derivation Assist, attorney's docket number 8222,

Application Serial No. --/ - - -, - - -, filed on same date herewith, by Brian D. Tate, entitled SQL-Based Automated, Adaptive, Histogram Bin Data Description Assist, attorney's docket number 8223,

Application Serial No. PCT/US99/ - - - - -, filed on same date herewith, by Timothy E. Miller, Brian D. Tate, Miriam H. Herman, Todd M. Brye, and Anthony L. Rollins, entitled Data Mining Assists in a Relational Database Management System, attorney's docket number 8224,

Application Serial No. --/ - - -, - - -, filed on same date herewith, by Todd M. Brye, Brian D. Tate, and Anthony L. Rollins, entitled SQL-Based Data Reduction Techniques for Delivering Data to Analytic Tools, attorney's docket number 8225,

Application Serial No. PCT/US99/ - - - - -, filed on same date herewith, by Timothy E. Miller, Miriam H. Herman, and Anthony L. Rollins, entitled Techniques for Deploying Analytic Models in Parallel, attorney's docket number 8226, and

Application Serial No. PCT/US99/- - - - -, filed on same date herewith, by Timothy E. Miller, Brian D. Tate, and Anthony L. Rollins, entitled Analytic Logical Data Model, attorney's docket number 8227, all of which are incorporated by reference herein.

BACKGROUND OF THE INVENTION

1. Field of the Invention.

This invention relates in general to a relational database management system, and in particular, to SQL-based analytic algorithms that provide statistical and machine learning methods to create analytic models from the data residing in a relational database.

2. Description of Related Art.

Relational databases are the predominate form of database management systems used in computer systems. Relational database management systems are often used in so-called "data warehouse" applications where enormous amounts of data are stored and processed. In recent years, several trends have converged to create a new class of data warehousing applications known as data mining applications. Data mining is the process of identifying and interpreting patterns in databases, and can be generalized into three stages.

Stage one is the reporting stage, which analyzes the data to determine what happened. Generally, most data warehouse implementations start with a focused application in a specific functional area of the business. These applications usually focus on reporting historical snap shots of business information that was previously difficult or impossible to access. Examples include Sales Revenue Reporting, Production Reporting and Inventory Reporting to name a few.

Stage two is the analyzing stage, which analyzes the data to determine why it happened. As stage one end-users gain previously unseen views of their business, they quickly seek to understand why certain events occurred; for example a decline in sales revenue. After discovering a reported decline in sales, data warehouse users will then obviously ask, "Why did sales go down?" Learning the answer to this

question typically involves probing the database through an iterative series of ad hoc or multidimensional queries until the root cause of the condition is discovered. Examples include Sales Analysis, Inventory Analysis or Production Analysis.

Stage three is the predicting stage, which tries to determine what will happen. As stage two users become more sophisticated, they begin to extend their analysis to include prediction of unknown events. For example, "Which end-users are likely to buy a particular product", or "Who is at risk of leaving for the competition?" It is difficult for humans to see or interpret subtle relationships in data, hence as data warehouse users evolve to sophisticated predictive analysis they soon reach the limits of traditional query and reporting tools. Data mining helps end-users break through these limitations by leveraging intelligent software tools to shift some of the analysis burden from the human to the machine, enabling the discovery of relationships that were previously unknown.

Many data mining technologies are available, from single algorithm solutions to complete tool suites. Most of these technologies, however, are used in a desktop environment where little data is captured and maintained. Therefore, most data mining tools are used to analyze small data samples, which were gathered from various sources into proprietary data structures or flat files. On the other hand, organizations are beginning to amass very large databases and end-users are asking more complex questions requiring access to these large databases.

Unfortunately, most data mining technologies cannot be used with large volumes of data. Further, most analytical techniques used in data mining are algorithmic-based rather than data-driven, and as such, there are currently little synergy between data mining and data warehouses. Moreover, from a usability perspective, traditional data mining techniques are too complex for use by database administrators and application programmers, and are too difficult to change for a different industry or a different customer.

Thus, there is a need in the art for data mining applications that directly operate against data warehouses, and that allow non-statisticians to benefit from advanced mathematical techniques available in a relational environment.

SUMMARY OF THE INVENTION

To overcome the limitations in the prior art described above, and to overcome other limitations that will become apparent upon reading and understanding the present specification, the present invention discloses a method,

apparatus, and article of manufacture for performing data mining applications in a relational database management system. At least one analytic algorithm is performed by a computer directly against a relational database, wherein the analytic algorithm includes SQL statements performed by the relational database management system and optional programmatic iteration, and the analytic algorithm creates at least one analytic model within an analytic logical data model from data residing in the relational database.

An object of the present invention is to provide more efficient usage of parallel processor computer systems. An object of the present invention is to provide a foundation for data mining tool sets in relational database management systems. Further, an object of the present invention is to allow data mining of large databases.

BRIEF DESCRIPTION OF THE DRAWINGS

Referring now to the drawings in which like reference numbers represent corresponding parts throughout:

FIG. 1 is a block diagram that illustrates an exemplary computer hardware environment that could be used with the preferred embodiment of the present invention;

FIG. 2 is a block diagram that illustrates an exemplary logical architecture that could be used with the preferred embodiment of the present invention; and

FIGS. 3, 4, and 5 are flowcharts that illustrate exemplary logic performed according to the preferred embodiment of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

In the following description of the preferred embodiment, reference is made to the accompanying drawings, which form a part hereof, and in which is shown by way of illustration a specific embodiment in which the invention may be practiced. It is to be understood that other embodiments may be utilized and structural changes may be made without departing from the scope of the present invention.

OVERVIEW

The present invention provides a relational database management system (RDBMS) that supports data mining operations of relational databases. In essence,

advanced analytic processing capabilities for data mining applications are placed where they belong, i.e., close to the data. Moreover, the results of these analytic processing capabilities can be made to persist within the database or can be exported from the database. These analytic processing capabilities and their results are exposed externally to the RDBMS by an application programmable interface (API).

According to the preferred embodiment, the data mining process is an iterative approach referred to as a "Knowledge Discovery Analytic Process" (KDAP). There are six major tasks within the KDAP:

1. Understanding the business objective.
2. Understanding the source data available.
3. Selecting the data set and "pre-processing" the data.
4. Designing the analytic model.
5. Creating and testing the models.
6. Deploying the analytic models.

The present invention provides various components for addressing these tasks:

- An RDBMS that executes Structured Query Language (SQL) statements against a relational database.
- An analytic Application Programming Interface (API) that creates scalable data mining functions comprised of complex SQL statements.
- Application programs that instantiate and parameterize the analytic API.
- Analytic algorithms utilizing:
  - Extended ANSI SQL statements,
  - a Call Level Interface (CLI) comprised of SQL staterments and programmatic iteration, and
  - a Data Reduction Utility Program comprised of SQL statements and programmatic iteration.
- An analytical logical data model (LDM) that stores results from and information about the advanced analytic processing in the RDBMS.
- A parallel deployer that controls parallel execution of the results of the analytic algorithms that are stored in the analytic logical data model.

The benefits of the present invention include:

- Data mining of very large databases directly within a relational database.
- Management of analytic results within a relational database.
- A comprehensive set of analytic operations that operate within a relational database management system.
- Application integration through an object-oriented API.

These components and benefits are described in more detail below.

HARDWARE ENVIRONMENT

FIG. 1 is a block diagram that illustrates an exemplary computer hardware environment that could be used with the preferred embodiment of the present invention. In the exemplary computer hardware environment, a massively parallel processing (MPP) computer system 100 is comprised of one or more processors or nodes 102 interconnected by a network 104. Each of the nodes 102 is comprised of one or more processors, random access memory (RAM), read-only memory (ROM), and other components. It is envisioned that attached to the nodes 102 may be one or more fixed and/or removable data storage units (DSUs) 106 and one or more data communications units (DCUs) 108, as is well known in the art.

Each of the nodes 102 executes one or more computer programs, such as a Data Mining Application (APPL) 110 performing data mining operations, Advanced Analytic Processing Components (AAPC) 112 for providing advanced analytic processing capabilities for the data mining operations, and/or a Relational Database Management System (RDBMS) 114 for managing a relational database 116 stored on one or more of the DSUs 106 for use in the data mining applications, wherein various operations are performed in the APPL 110, AAPC 112, and/or RDBMS 114 in response to commands from one or more Clients 118. In alternative embodiments, the APPL 110 may be executed in one or more of the Clients 118, or on an application server on a different platform attached to the network 104.

Generally, the computer programs are tangibly embodied in and/or retrieved from RAM, ROM, one or more of the DSUs 106, and/or a remote device coupled to the computer system 100 via one or more of the DCUs 108. The computer programs comprise instructions which, when read and executed by a

node 102, causes the node 102 to perform the steps necessary to execute the steps or elements of the present invention.

Those skilled in the art will recognize that the exemplary environment illustrated in FIG. 1 is not intended to limit the present invention. Indeed, those skilled in the art will recognize that other alternative hardware environments may be used without departing from the scope of the present invention. In addition, it should be understood that the present invention may also apply to other computer programs than those disclosed herein.

LOGICAL ARCHITECTURE

FIG. 2 is a block diagram that illustrates an exemplary logical architecture of the AAPC 112, and its interaction with the APPL 110, RDBMS 114, relational database 116, and Client 118, according to the preferred embodiment of the present invention. In the preferred embodiment, the AAPC 112 includes the following components:

- An Analytic Logical Data Model (LDM) 200 that stores results from the advanced analytic processing in the RDBMS 114,
- One or more Scalable Data Mining Functions 202 that comprise complex, optimized SQL statements that perform advanced analytic processing in the RDBMS 114,
- An Analytic Application Programming Interface (API) 204 that provides a mechanism for an APPL 110 or other component to invoke the Scalable Data Mining Functions 202,
- One or more Analytic Algorithms 206 that can operate as standalone applications or can be invoked by another component, wherein the Analytic Algorithms 206 comprise:
  - Extended ANSI SQL 208 that can be used to implement a certain class of Analytic Algorithms 206,
  - A Call Level Interface (CLI) 210 that can be used when a combination of SQL and programmatic iteration is required to implement a certain class of Analytic Algorithms 206, and
  - A Data Reduction Utility Program 212 that can be used to implement a certain class of Analytic Algorithms 206 where data is first reduced using SQL followed by programmatic iteration.

- An Analytic Algorithm Application Programming Interface (API) 214 that provides a mechanism for an APPL 110 or other components to invoke the Analytic Algorithms 206,
- A Parallel Deployer 216 that controls parallel executions of the results of an Analytic Algorithm 206 (sometimes referred to as an analytic model) that are stored in the Analytic LDM 200, wherein the results of executing the Parallel Deployer 216 are stored in the RDBMS 114.

Note that the use of these various components is optional, and thus only some of the components may be used in any particular configuration.

The preferred embodiment is oriented towards a multi-tier logical architecture, in which a Client 118 interacts with the various components described above, which, in turn, interface to the RDBMS 114 to utilize a large central repository of enterprise data stored in the relational database 116 for analytic processing.

In one example, a Client 118 interacts with an APPL 110, which interfaces to the Analytic API 204 to invoke one or more of the Scalable Data Mining Functions 202, which are executed by the RDBMS 114. The results from the execution of the Scalable Data Mining Functions 202 would be stored as an analytic model within an Analytic LDM 200 in the RDBMS 114.

In another example, a Client 118 interacts with one or more Analytic Algorithms 206 either directly or via the Analytic Algorithm API 214. The Analytic Algorithms 206 comprise SQL statements that may or may not include programmatic iteration, and the SQL statements are executed by the RDBMS 114. In addition, the Analytic Algorithms 206 may or may not interface to the Analytic API 204 to invoke one or more of the Scalable Data Mining Functions 202, which are executed by the RDBMS 114. Regardless, the results from the execution of the Analytic Algorithms 206 would be stored as an analytic model within an Analytic LDM 200 in the RDBMS 114.

In yet another example, a Client 118 interacts with the Parallel Deployer 216, which invokes parallel instances of the results of the Analytic Algorithms 206, sometimes referred to as an Analytic Model. The Analytic Model is stored in the Analytic LDM 200 as a result of executing an instance of the Analytic Algorithms 206. The results of executing the Parallel Deployer 216 are stored in the RDBMS 114.

In still another example, a Client 118 interacts with the APPL 110, which invokes one or more Analytic Algorithms 206 either directly or via the Analytic Algorithm API 214. The results would be stored as an analytic model within an Analytic LDM 200 in the RDBMS 114.

The overall goal is to significantly improve the performance, efficiency, and scalability of data mining operations by performing compute and/or I/O intensive operations in the various components. The preferred embodiment achieves this not only through the parallelism provided by the MPP computer system 100, but also from reducing the amount of data that flows between the APPL 110, AAPC 112, RDBMS 114, Client 118, and other components.

Those skilled in the art will recognize that the exemplary configurations illustrated and discussed in conjunction with FIG. 2 are not intended to limit the present invention. Indeed, those skilled in the art will recognize that other alternative configurations may be used without departing from the scope of the present invention. In addition, it should be understood that the present invention may also apply to other components than those disclosed herein.

Scalable Data Mining Functions

The Scalable Data Mining Functions 202 comprise complex, optimized SQL statements that are created, in the preferred embodiment, by parameterizing and instantiating the corresponding Analytic APIs 204. The Scalable Data Mining Functions 202 perform much of the advanced analytic processing for data mining applications, when performed by the RDBMS 114, without having to move data from the relational database 116.

The Scalable Data Mining Functions 202 can be categorized by the following functions:

- Data Description: The ability to understand and describe the available data using statistical techniques. For example, the generation of descriptive statistics, frequencies and/or histogram bins.
- Data Derivation: The ability to generate new variables (transformations) based upon existing detailed data when designing an analytic model. For example, the generation of predictive variables such as bitmaps, ranges, codes and mathematical functions.

- Data Reduction: The ability to reduce the number of variables (columns) or observations (rows) used when designing an analytic model. For example, creating Covariance, Correlation, or Sum of Squares and Cross-Products (SSCP) Matrices.
- Data Reorganization: The ability to join or denormalize pre-processed results into a wide analytic data set.
- Data Sampling/Partitioning: The ability to intelligently request different data samples or data partitions. For example, hash data partitioning or data sampling.

The principal theme of the Scalable Data Mining Functions 202 is to facilitate analytic operations within the RDBMS 114, which process data collections stored in the database 116 and produce results that also are stored in the database 116. Since data mining operations tend to be iterative and exploratory, the database 116 in the preferred embodiment comprises a combined storage and work space environment. As such, a sequence of data mining operations is viewed as a set of steps that start with some collection of tables in the database 116, generate a series of intermediate work tables, and finally produce a result table or view.

Analytic Algorithms

The Analytic Algorithms 206 provide statistical and "machine learning" methods to create Analytic LDMs 200 from the data residing in the relational database 116. Analytic Algorithms 206 that are completely data driven, such as association, can be implemented solely in Extended ANSI SQL 208. Analytic Algorithms 206 that require a combination of SQL and programmatic iteration, such as induction, can be implemented using the CLI 210. Finally, Analytic Algorithms 206 that require almost complete programmatic iteration, such as clustering, can be implemented using a Data Reduction Utility Program 212, wherein this approach involves data pre-processing that reduces the amount of data that a non-SQL algorithm can then process.

The Analytic Algorithms 206 significantly improve the performance and efficiency of data mining operations by providing the technology components to perform advanced analytic operations directly against the RDBMS 114. In addition, the Analytic Algorithms 206 leverage the parallelism that exists in the MPP computer system 100, the RDBMS 114, and the database 116.

The Analytic Algorithms 206 provide data analysts with an unprecedented option to train and apply "machine learning" analytics against massive amounts of data in the relational database 116. Prior techniques have failed as their sequential design is not optimal in an RDBMS 114 environment. Because the Analytic Algorithms 206 are implemented in Extended ANSI SQL 208, through the CLI 210, and/or by means of the Data Reduction Utility Program 212, they can therefore leverage the scalability available on the MPP computer system 100. In addition, taking a data-driven approach to analysis, through the use of complete Extended ANSI SQL 208, allows people other than highly educated statisticians to leverage the advanced analytic techniques offered by the Analytic Algorithms 206.

Extended ANSI SQL

As mentioned above, Analytic Algorithms 206 that are completely data driven, such as affinity analysis, can be implemented solely in Extended ANSI SQL 208. Typically, these type of algorithms operate against a set of tables in the relational database 116 that are populated with transaction-level data, the source of which could be point-of-sale devices, automated teller machines, call centers, the Internet, etc. The SQL statements used to process this data typically build relationships between and among data elements in the tables. For example, the SQL statements used to process data from point-of-sale devices may build relationships between and among products and pairs of products. Additionally, the dimension of time can be added in such a way that these relationships can be analyzed to determine how they change over time. As the implementation is solely in SQL statements, the design takes advantage of the hardware and software environment of the preferred embodiment by decomposing the SQL statements into a plurality of sort and merge steps that can be executed concurrently in parallel by the MPP computer system 100.

Call-Level Interface

As mentioned above, Analytic Algorithms 206 that require a mix of programmatic iteration along with Extended ANSI SQL statements, such as inductive inference, can be implemented using the CLI 210. Whereas the SQL approach is appropriate for business problems that are descriptive in nature, inference problems are predictive in nature and typically require a training phase where the APPL 110 "learns" various rules based upon the data description,

followed by testing and application, and where the rules are validated and applied against a new data set. This class of algorithms are compute-intensive and historically can not handle large volumes of data because they expect the analyzed data to be in a specific fixed or variable flat file format.

Most implementations first extract the data from the database 116 to construct a flat file and then execute the "train" portion on this resultant file. This method is slow and limited by the amount of memory available in the computer system 100. This process can be improved by leveraging the relational database 116 to perform those portions of the analysis, instead of extracting all the data.

When SQL statements and programmatic iteration are used together, the RDBMS 114 can be leveraged to perform computations and order data within the relational database 116, and then extract the information using very little memory in the APPL 110. Additionally, computations, aggregations and/or ordering can be run in parallel, because of the massively parallel nature of the RDBMS 114.

Data Reduction Utility Program

As mentioned above, Analytic Algorithms 206 that can operate on a reduced or scaled data set, such as regression or clustering, the Data Reduction Utility Program 212 can be used. The problem of creating analytic models from massive amounts of detailed data has often been addressed by sampling, mainly because compute intensive algorithms cannot handle large volumes of data. The approach of the Data Reduction Utility Program 212 is to reduce data through operations such as matrix calculations or histogram binning, and then use this reduced or scaled data as input to a non-SQL algorithm. This method intentionally reduces fine numerical data details by assigning them to ranges, or bins, correlating their values or determining their covariances. The capacity of the preferred embodiment for creating these data structures from massive amounts of data in parallel gives it a special opportunity in this area.

Analytic Logical Data Model

The Analytic LDM 200, which is integrated with the relational database 116 and the RDBMS 114, provides logical entity and attribute definitions for advanced analytic processing, i.e., the Scalable Data Mining Functions 202 and Analytic Algorithms 206, performed by the RDBMS 114 directly against the relational database 116. These logical entity and attribute definitions comprise metadata that

define the characteristics of data stored in the relational database 116, as well as metadata that determines how the RDBMS 114 performs the advanced analytic processing. The Analytic LDM 200 also stores processing results from this advanced analytic processing, which includes both result tables and derived data for the Scalable Data Mining Functions 202, Analytic Algorithms 206, and the Parallel Deployer 216. The Analytic LDM 200 is a dynamic model, since the logical entities and attributes definitions change depending upon parameterization of the advanced analytic processing, and since the Analytic LDM 200 is updated with the results of the advanced analytic processing.

Logic of the Preferred Embodiment

Flowcharts which illustrate the logic of the preferred embodiment of the present invention are provided in FIGS. 3, 4 and 5. Those skilled in the art will recognize that this logic is provided for illustrative purposes only and that different logic may be used to accomplish the same results.

Referring to FIG. 3, this flowchart illustrates the logic of the Scalable Data Mining Functions 202 according to the preferred embodiment of the present invention.

Block 300 represents the one or more of the Scalable Data Mining Functions 202 being created via the API 204. This may entail, for example, the instantiation of an object providing the desired function.

Block 302 represents certain parameters being passed to the API 204, in order to control the operation of the Scalable Data Mining Functions 202.

Block 304 represents the metadata in the Analytic LDM 200 being accessed, if necessary for the operation of the Scalable Data Mining Function 202.

Block 306 represents the API 204 generating a Scalable Data Mining Function 204 in the form of a data mining query based on the passed parameters and optional metadata.

Block 308 represents the Scalable Data Mining Function 204 being passed to the RDBMS 114 for execution.

Referring to FIG. 4, this flowchart illustrates the logic of the Analytic Algorithms 206 according to the preferred embodiment of the present invention.

Block 400 represents the Analytic Algorithms 206 being invoked, either directly or via the Analytic Algorithm API 214.

Block 402 represents certain parameters being passed to the Analytic Algorithms 206, in order to control their operation.

Block 404 represents the metadata in the Analytic LDM 200 being accessed, if necessary for the operation of the Analytic Algorithms 206.

Block 406 represents the Analytic Algorithms 206 passing SQL statements to the RDBMS 114 for execution and Block 408 optionally represents the Analytic Algorithms 206 performing programmatic iteration. Those skilled in the art will recognize that the sequence of these steps may differ from those described above, may not include both steps, may include additional steps, and may include iterations of these steps.

Block 410 represents the Analytic Algorithms 206 storing results in the Analytic LDM 200.

Referring to FIG. 5, this flowchart illustrates the logic performed by the RDBMS 114 according to the preferred embodiment of the present invention.

Block 500 represents the RDBMS 114 receiving a query or other SQL statements.

Block 502 represents the RDBMS 114 analyzing the query.

Block 504 represents the RDBMS 114 generating a plan that enables the RDBMS 114 to retrieve the correct information from the relational database 116 to satisfy the query.

Block 506 represents the RDBMS 114 compiling the plan into object code for more efficient execution by the RDBMS 114, although it could be interpreted rather than compiled.

Block 508 represents the RDBMS 114 initiating execution of the plan.

Block 510 represents the RDBMS 114 generating results from the execution of the plan.

Block 512 represents the RDBMS 114 either storing the results in the Analytic LDM 200, or returning the results to the Analytic Algorithm 206, APPL 110, and/or Client 118.

CONCLUSION

This concludes the description of the preferred embodiment of the invention. The following describes an alternative embodiment for accomplishing the same invention. Specifically, in an alternative embodiment, any type of

computer, such as a mainframe, minicomputer, or personal computer, could be used to implement the present invention.

In summary, the present invention discloses a method, apparatus, and article of manufacture for performing data mining applications in a relational database management system. At least one analytic algorithm is performed by a computer directly against a relational database, wherein the analytic algorithm includes SQL statements performed by the relational database management system and optional programmatic iteration, and the analytic algorithm creates at least one analytic model within an analytic logical data model from data residing in the relational database.

The foregoing description of the preferred embodiment of the invention has been presented for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form disclosed. Many modifications and variations are possible in light of the above teaching. It is intended that the scope of the invention be limited not by this detailed description, but rather by the claims appended hereto.

WHAT IS CLAIMED IS:

1. A computer-implemented system for performing data mining applications, comprising:

(a) a computer having one or more data storage devices connected thereto;

(b) a relational database management system, executed by the computer, for managing a relational database stored on the data storage devices; and

(c) at least one analytic algorithm performed by the computer, wherein the analytic algorithm includes SQL statements performed by the relational database management system directly against the relational database and optional programmatic iteration, and the analytic algorithm creates at least one analytic model within an analytic logical data model from data residing in the relational database.

2. The computer-implemented system of claim 1, wherein the analytic algorithm provides statistical and machine learning methods for creating the analytic logical data model.

3. The computer-implemented system of claim 1, wherein the analytic algorithm is implemented in Extended ANSI SQL.

4. The computer-implemented system of claim 3, wherein the analytic algorithm operates against a set of tables in the relational database, and the Extended ANSI SQL build relationships among data elements in the tables.

5. The computer-implemented system of claim 4, wherein the Extended ANSI SQL analyzes the relationships to determine how the relationships change..

6. The computer-implemented system of claim 1, wherein the analytic algorithm is implemented in a Call Level Interface (CLI) that processes data from the relational database using SQL and programmatic iteration.

7. The computer-implemented system of claim 6, wherein the CLI is used with SQL to perform computations, aggregations, and/or ordering on the data from the relational database.

8. The computer-implemented system of claim 1, wherein the analytic algorithm is implemented by a Data ReductionUtility Program that reduces data from the relational database in bulk using SQL followed by a non-SQL iterative program..

9. The computer-implemented system of claim 8, wherein the Data Reduction Utility Program provides a sequence of Extended ANSI SQL followed by programmatic iteration.

10. A method for performing data mining applications, comprising:

(a) managing a relational database stored on one or more data storage devices connected to a computer; and

(b) performing at least one analytic algorithm in the computer, wherein the analytic algorithm includes SQL statements performed by a relational database management system directly against the relational database and optional programmatic iteration, and the analytic algorithm creates at least one analytic model within an analytic logical data model from data residing in the relational database.

11. An article of manufacture comprising logic embodying a method for performing data mining applications, comprising:

(a) managing a relational database stored on one or more data storage devices connected to a computer; and

(b) performing at least one analytic algorithm in the computer, wherein the analytic algorithm includes SQL statements performed by a relational database management system directly against the relational database and optional programmatic iteration, and the analytic algorithm creates at least one analytic model within an analytic logical data model from data residing in the relational database.

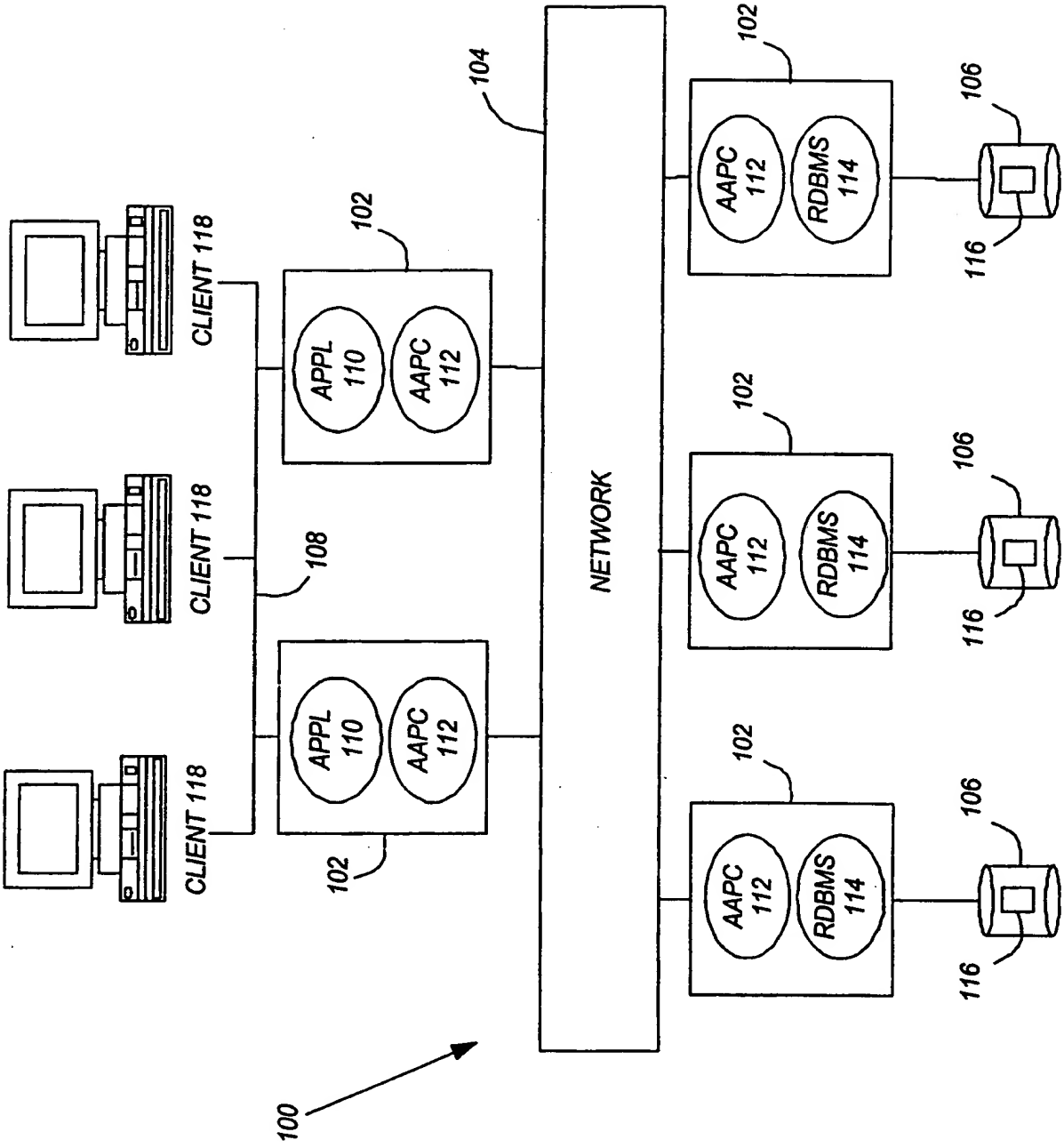

100
NETWORK
104
CLIENT 118
CLIENT 118
CLIENT 118
108
APPL 110
AAPC 112
102
AAPC 112
RDBMS 114
106
116

FIG. 1

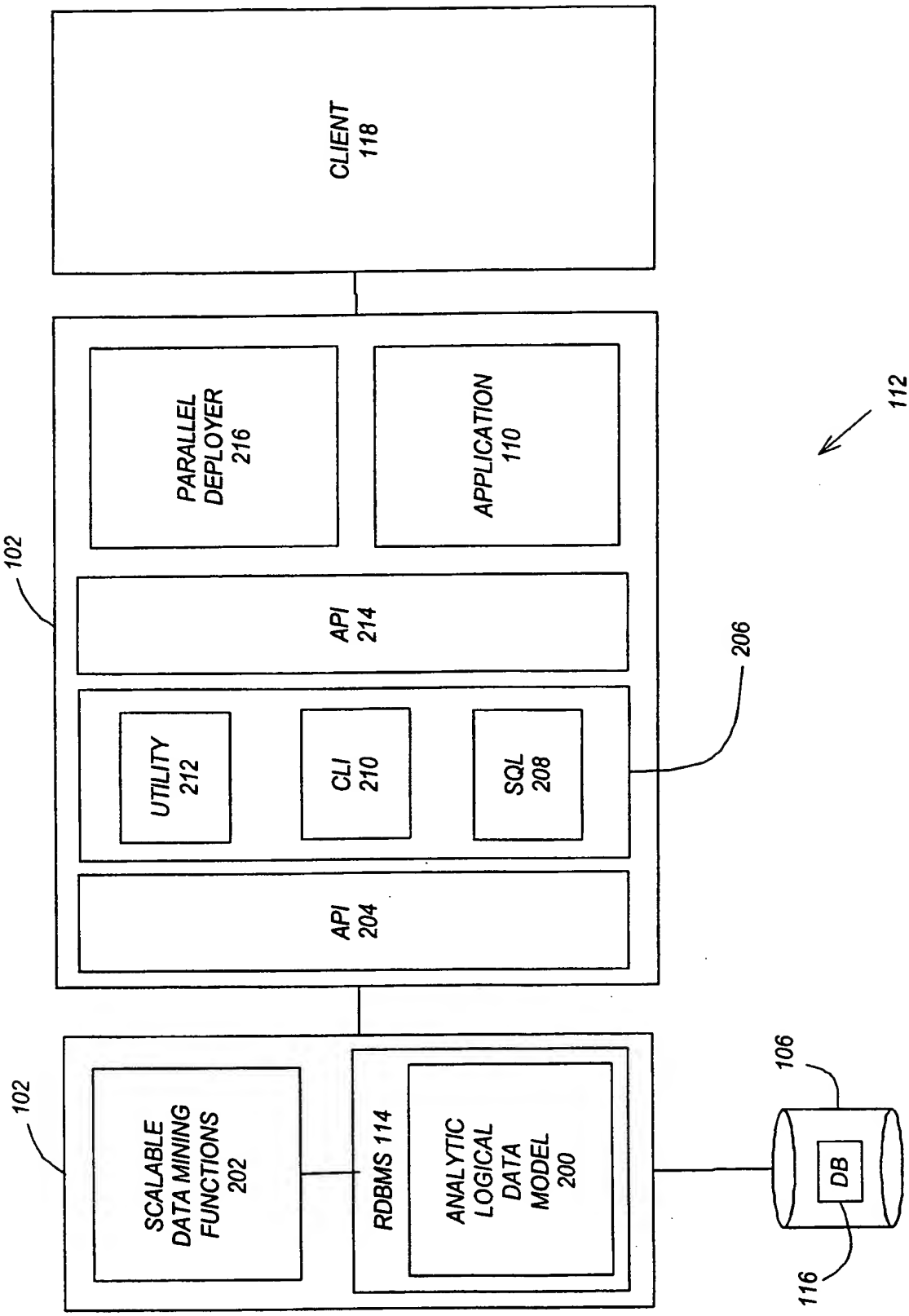

CLIENT
118
PARALLEL
DEPLOYER
216
APPLICATION
110
112
102
API
214
206
UTILITY
212
CLI
210
SQL
208
API
204
102
SCALABLE
DATA MINING
FUNCTIONS
202
RDBMS 114
ANALYTIC
LOGICAL
DATA
MODEL
200
106
DB
116

FIG. 2

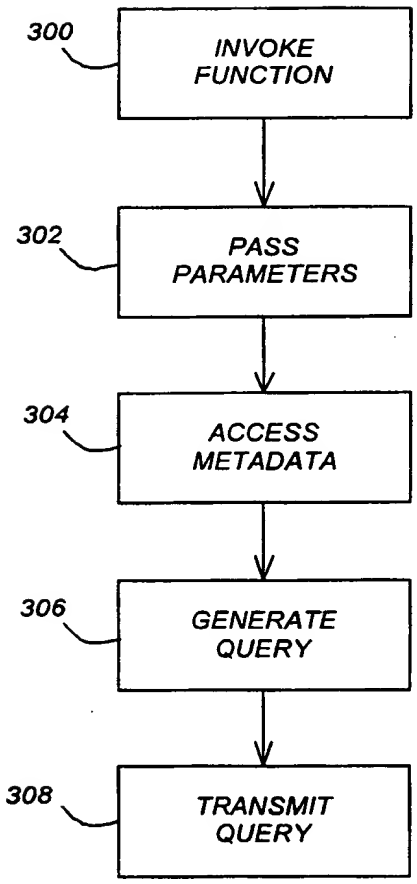

300
INVOKE FUNCTION
302
PASS PARAMETERS
304
ACCESS METADATA
306
GENERATE QUERY
308
TRANSMIT QUERY

FIG. 3

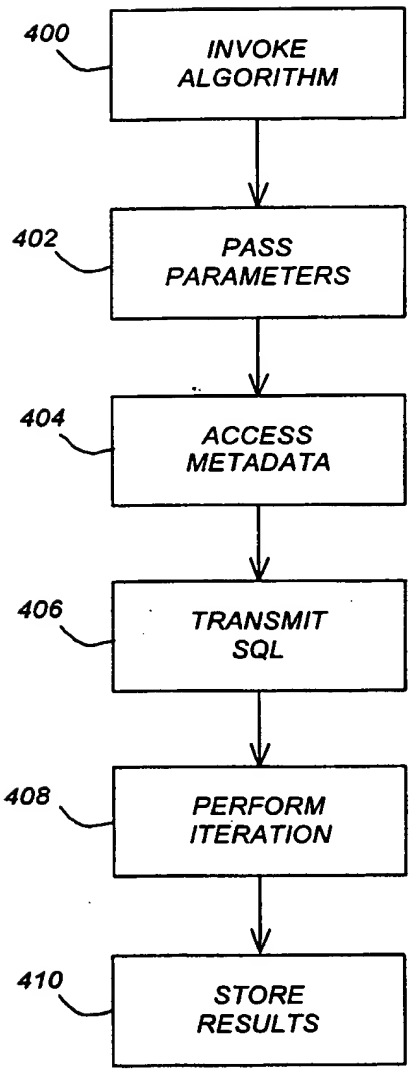

400
INVOKE ALGORITHM
402
PASS PARAMETERS
404
ACCESS METADATA
406
TRANSMIT SQL
408
PERFORM ITERATION
410
STORE RESULTS

FIG. 4

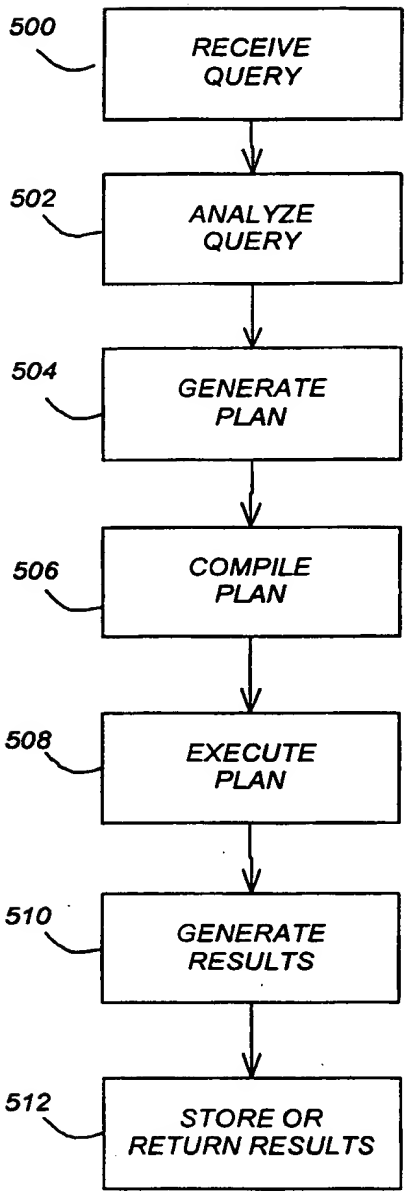

500
RECEIVE QUERY
502
ANALYZE QUERY
504
GENERATE PLAN
506
COMPILE PLAN
508
EXECUTE PLAN
510
GENERATE RESULTS
512
STORE OR RETURN RESULTS

FIG. 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US99/22966

A. CLASSIFICATION OF SUBJECT MATTER
IPC(6) :G06F 15/16, 17/00, 17/30
US CL :707/2, 4, 100, 102
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED
Minimum documentation searched (classification system followed by classification symbols)
U.S. : 707/2, 4, 100, 102

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
NONE

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
WEST, EAST

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5, 412,806 A (DU et al.) 02 May 1995, the entire paper is relevant	1-11
Y	US 5,590,322 A (HARDING et al.) 31 December 1996, the entire paper is relevant	1-11
Y	US 5,799,310 A (ANDERSON et al.) 25 August 1998, the entire paper is relevant	1-11
Y	US 5, 806,066 A (GOLSHANI et al.) 08 September 1998, the entire paper is relevant	1-11

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search
08 DECEMBER 1999

Date of mailing of the international search report
23 DEC 1999

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231
Facsimile No. (703) 305-3230

Authorized officer
THUY PARDO
Telephone No. (703) 305-1091